VCDE Silver Level Interoperability Review
caTIES 1.1

December 1, 2005

Companion Document:
"caTIES Interoperability Review 20051201.xls"

Review Team:
David Aronow (Lead)
Brian Davis
Robert Freimuth
Lewis Frey
Michael Keller
Rakesh Nagarajan
Hemant Shah

INTRODUCTION

Purpose

The purpose of this document is to provide the Vocabulary and Common Data Elements (VCDE) working group with a summary description of quantitative and qualitative criteria for judging if the quality of the information in the caTIES 1.1 Interoperability Submission Package is sufficient for syntactically and semantically interoperablity within caBIG at the silver level. The guidance and criteria contained herein is intended to report the evaluation of the caTIES system and to inform the designs of new systems. A format for organizing and communicating the evaluation of the information in the Interoperability Submission Package is presented in the accompanying Interoperability Review excel spreadsheet.

Interoperability can be defined as the ability of a system to access and use the parts of another system. The problem of access is a problem of poor syntactic interoperability. Regularization of application programming and messaging interfaces is necessary to overcome obstacles to syntactic interoperability. The problem of usage is a problem of poor or ambiguous semantic interoperability. Explicit descriptions and definitions of the contents and meanings of resources are necessary to overcome barriers to semantic interoperability at the silver level of maturity. Silver level maturity is defined as

> **A rigorous set of requirements that, when met, significantly reduce the barrier to use of a resource by a remote party who was not involved in the development of that resource.** This level requires substantial interaction with the caBIG cross-cutting Workspaces (the Vocabularies and Common Data Elements Workspace or the Architecture Workspace).

The Interoperability Review is at the attribute level specifically examining identifiers that can be potentially shared between classes. So, for example, if two classes both share deidentifiedId as an attribute, a distributed join on the objects can be performed which would come up with a *meaningful* superset. To support interoperability at the level of joining classes, key attributes need to be identified and thoroughly reviewed.

EXECUTIVE SUMMARY

The review team appreciates that the 1.1 release of caTIES has already been implemented in caGrid 0.5. None of the findings, requirements or recommendations of this review are intended to impact that reference implementation.

The review team concludes that the 1.1 release of caTIES is Conditionally Acceptable at the caBIG Silver Level. The review team used color and type characteristics to indicate compliance going forward.

- ***Red: Issues, underlined and italicized,*** that are required to be corrected in the next release of caTIES.
- Yellow: ***Italicized points of concern***, including recommendations to be considered by the caTIES developers and open issues that need further discussion.
- Green: Satisfactory compliance has been achieved.

The companion Excel Spreadsheet organizes the material used for the review into the following topics: Description, Key, Comment, Association, View and SIW Error.

- Description lists for each class: the class name, class description, attributes and attribute descriptions.
- Key provides a list of keys that can potentially be used for joining across class. For each class, it contains a list of the class name, comments, keys and definitions.
- Comments are observations and questions that have been documented by the interoperability team members.
- Associations are the association relationships between the classes in a matrix format. Red indicates missing directionality for associations.
- View corresponds to issues related to the views into the UML model.
- SIW Error provides the error report for the Semantic Integration Workbench (SIW).

The following compliance section goes over the details of the review and the accompanying spreadsheet organizes complex information used to complete the review.

COMPATIBILITY

When considering how to overcome the obstacles to interoperability, the caBIG program members arrived at four areas that need to be addressed. Three of the four areas address issues related to semantic interoperability and the remaining one addresses issues related to syntactic interoperability. The four areas are: Common Data Elements, Information Models, Vocabulary & Ontologies, and Programming & Messaging Interfaces. Each of these will be examined for the Silver Level Interoperability Review of caTIES.

Data Elements

Common Data Elements. Data that is collected on a given study or trial must be defined and described such that remote users of that data can understand what it means. These metadata descriptions are referred to as data elements. When many groups use the same [common] data elements (CDEs), then larger-scale studies can be conceived, since consistency and comparability across sites, studies, and time becomes possible. CDEs are therefore critical constructs for semantic interoperability.

CDE Bronze-Level Prerequisites for Silver Level:

1) Data elements used in APIs are identified

   We need a link to the documention for the APIs. This is needed to validate if the elements used in the API are identified.

   A web services description was given at the following link:

   http://caties.cabig.upmc.edu/ogsa/services/cagrid/caTIES?wsdl. We were unable to validate this point with the above link.

2) Data Elements have names, definitions, datatypes, and permissible values with meanings suitable for unambiguous interpretation.

   This is validated by a clean error report from SIW other than permissible values.

3) Data Elements are built using a controlled vocabulary.

   Vocabularies used are well established vocabularies.

4) Data Element metadata is in an electronic format

   This is validated by a clean error report from SIW on the annotated xmi file.

5) Data Element metadata is exposed in a publicly accessible electronic resource that is distinct from the information system itself.

   This is validated by a clean error report from SIW and error free load into caDSR.

CDE Silver Level Requirements:

CDEs for Race, Ethnicity, Gender, DateTime and Date use caDSR identifiers, but do not include all components of those, Standard, CDEs. The caTIES development team reports that it intends "to implement some standards in next version," following a strategy of "use DECs identical to standard for now, and map to the standard VD (and CDE) later."

Standard CDEs are required to be implemented for all appropriate caTIES CDEs in full in the next release of caTIES. The following compliance assessment assumes this will be done.

1) These CDEs are all registered in the caBIG Context of the NCI cancer Data Standards Repository (caDSR), an implementation of the ISO/IEC11179 standard for metadata registries.

   This is validated via the SIW error report.

2) Reuse of existing validated CDEs in the caDSR must be considered before any new data elements are created. All new CDEs are subject to review and validation by the VCDE workspace before they are deployed. Existing CDEs are reused (the public ID of a reused CDE is recorded in the metadata). Reuse percentage can be provided.

   All CDE currently treated as not reused, but having Workflow Status of Draft New. Conformance is expected in next release.

3) Administered Components are built on terminologies that have been reviewed and validated by the caBIG VCDE Workspace. Vocabularies are registered in EVS.

   They will be in conformance in the next release.

4) All reused Data Elements have been designated.

   All CDE currently treated as not reused, but are in caBIG context.

5) All Administered Components are constructed according to best practices defined by the caBIG VCDE workspace

   See Comments page of Excel spreadsheet for specific concerns.

   a. All Data Element Concepts (DEC) have object and properties assigned from EVS.

      This is validated by a clean error report from SIW.

   b. All Value Domains (VD) have a representation assigned from EVS.

      This is validated by a clean error report from SIW.

   c. All Data Elements (DE) are associated with a unique pairing of DEC and VD.

      This is validated by a clean error report from SIW.

   d. All DECs are associated with a Conceptual Domain (CD).

      This is validated by a clean error report from SIW.

   e. Permissible values are documented with value meanings and are related to EVS terms.

      i. Expected with implementation of Standard CDES, discussed above.

      ii. CDEs "Concept Reference Object Qualifier Flag" and "Concept Reference Object Negation Flag" have Datatype = Boolean and Type = Non Enumerated. Type should be changed to Enumerated with permissible values provided in the next release.

iii. CDEs "NCI Metathesaurus Code" and "Negation NCI Metathesaurus Code" have codes as values. Are these codes biomedical concepts from EVS?

f. All Administered Components have names and definitions.

CDEs "PathologyReport NCI Metathesaurus Code" and "PathologyReport Negation NCI Metathesaurus Code" need defintions that specify they are concatenated strings of codes and the method of delimitation.

g. All Value Domains have datatypes.

This is validated by a clean error report from SIW as long as there are valid datatypes.

h. All Administered Components have a Workflow Status of Released.

Workflow status is DRAFT NEW.

Regarding correct use of Administered Components:

CDE "Patient Identifier Deidentification" has the Property and PropertyQualifier1 reversed. As in "Pathology Report Deidentification Identifier", the Property = "Identifier" and "Deidentification" is but one of the Qualifiers for that Property that is used in this submission.

CDEs "PathologyReport Document Text" and "PathologyReport Document Extensible Markup Language" have the Property and PropertyQualifier1 reversed. As in "Binary Document", the Property = "Document" and "Text" or "Extensible Markup Language" are but two of the Qualifiers for that Property that is used in this submission.

6) New CDEs and related metadata have been reviewed and validated by VCDE workspace.

Per this review report.

Information Models

Data Elements are precise specifications of individual types of data that are collected during a research study or using measurement technologies. However, scientific interpretation relies on the placement of data elements into a broader semantic context, an information model. Therefore, in order to attain the highest degree of semantic interoperability, data must be expressed in the context of such a model.

Information Models. Individual types of data are rarely collected or presented in isolation. Rather, they are assembled into a contextual environment that includes closely

and more distantly associated data and information. These associations and relationships can be presented in the form of an information model. These models convey both a human and a machine readable representation of the contextual environment of data in an information resource, and are important for achieving the highest degree semantic interoperability.

Silver-level compatibility requires

See Comments page of Excel spreadsheet for specific concerns.

1) The use of the industry-standard modeling language, UML, to create domain models that describe the content of the system.

   caTIES is represented with a UML model.

2) UML class diagrams that illustrate the data classes, attributes, and relationships are required. (Using other aspects of UML modeling is encouraged as a best practice in development methodology, but is not central to the issue of semantic interoperability).

   caTIES is represented with at UML model containing classes, attributes and associative relationships.

3) Class diagrams must conform to the naming conventions and terminology standards prescribed by the caBIG program. Use Package, Class and Attribute names that convey meaning clearly. Use Sound Object –oriented analysis and design techniques

   a. All Classes must be represented by Objects that are defined in a vocabulary (presently EVS)

      This is validated by a clean error report from SIW.

   b. Data Element Concepts must be represented by Class/Attribute Pairs

      This is validated by a clean error report from SIW.

   c. Association ends must be given a meaningful role name. Naming convention, Use associated class name. If multiplicity > 1, append "Collection"

      See Comments page of Excel spreadsheet for specific concerns.

   d. Association End Multiplicity must be explicitly defined

      Reviewer validated.

   e. Direction of association must be explicitly defined (as it determines visibility of objects in API)

      Directions of association are not defined in the model.

   f. ALL classes and attributes need description

All classes and attributes are described.

g. Elements must reside in "Logical Model" package

Reviewer validated.

h. Attribute types must be java primitive types

Attribute types are loaded consistently.

Currently the UML loader enforces java wrappers around the java primitive types.

i. UpperCamelCase for ClassName

Reviewer validated.

j. lowerCamelCase for attributeName

Reviewer validated.

k. UNDERLINED_CAPS for CONSTANT_NAME

Reviewer validated.

l. Avoid using class identifiers in attribute names (eg Use id, name instead of geneId, chromosomeName)

Reviewer validated.

m. Avoid using abbreviations and acronyms

Reviewer validated.

n. Avoid technical jargon

Reviewer validated.

o. Class and attribute names are ALWAYS singular

PathologyReport conceptCodes & negatedConceptCodes are plural.

4) UML models must be fully annotated with class and attribute definitions, and with associated terminology concept codes.

Other than item "a", these are validated by a clean error report from SIW.

a. UML Entity has outside, concrete meaning (is not abstract)

Reviewer validated.

b. UML Class = ISO Object Class in Terminology

c. UML Class Attribute = ISO Property in Terminology

d. UML Class + UML Class Attribute = ISO Data Element Concept (DEC) in CDE

e. UML Java Datatype = Iso Value Domain

f. UML Class + UML Class Attribute +Datatype/Valid Values = ISO CDE

5) The models must be provided in XML Metadata Interchange (XMI) format in addition to any diagrammatic representations (in GME).

Other than items "d & e", these are validated by a clean error report from SIW for the annotated xmi file.

a. XMI file in XMI 1.1, according to UML 1.3

b. Unisys/Rose Extensions

c. Export Tagged values

d. Include "logical model" package

Reviewer validated.

e. XMI file stored in (GME)

6) Upon review and validation by the VCDE workspace, models must be submitted for registration and loading into the caDSR.

Vocabularies & Ontologies

Vocabularies and Ontologies. Biomedical information includes a substantial body of specialized concepts that are represented by terms. Agreement upon the basic concepts, terms and definitions that are inherent in all biomedical information is essential for achieving semantic interoperability. Terminology development systems that use description logic are helpful tools for managing these concepts.

Silver-level maturity introduces the requirement for review and approval of terminologies by the caBIG VCDE workspace.

See Comments page of Excel spreadsheet for specific concerns.

Programming & Messaging Interfaces

Programming and Messaging Interfaces. Computer programs and the people who write them are able to access resources from other programs through programming and messaging interfaces. Each of these interfaces responds to a particular syntax for its communications. Agreement upon standards for these interfaces is necessary to overcome barriers to syntactic interoperability. One of these is an Application Programming Interface (API).

Silver-level compatibility is more demanding.

1) Data-oriented systems must provide a well-documented public API that is based upon an object-oriented abstraction of the underlying data. This abstraction layer

must be derived from a domain information model constructed in the Unified Modeling Language (UML; see *Information Models* below)

2) Data must be returned in the form of data objects that are instances of classes in the model. Data formats must conform to standards set by the caBIG workspace with which the resource is aligned.

3) Wherever use cases indicate a messaging system is warranted, a standards-based messaging protocol approved by the caBIG Architecture Workspace is used to exchange information. Silver-level analytical tools and client applications must be able to read directly from these caBIG-compatible interfaces.

The documentation for the APIs should be made available via a URL.